修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院　情報システム　学研究科　情報システム基盤学専攻　博士前期課程		
氏　　名	岡　遼太郎	学籍番号	0753004
論 文 題 目	単語間の類似性に基づく特徴ベクトルとカテゴリ別辞書を用いた文書識別		

要　　旨

コンピュータやインターネットの普及により,情報の電子化,また特定の人間のみが発信できるのではなく,ユーザである全ての人が情報を発信可能となったため,増大する情報(情報爆発)は人間の処理能力をはるかに超え,情報すべてに注意を払うことはもはや不可能である.そのため,情報を効率的に利用するため文書の内容に基づいて分類する必要性が増している.

そこで、本論文では,単語間の類似性に基づく特徴ベクトルとカテゴリ別辞書を用いた文書識別法を提案する.提案法は,従来の Term Frequency Inverse Document(TF-IDF 法)のような単語頻度ヒストグラムに基づいて作成した特徴ベクトルを用いた識別手法に対し,以下の３つの新規性を持つ手法の提案である.

（１） カテゴリ内の単語出現頻度(TF),カテゴリ内の同一単語が現れる文書頻度(DF),およびカテゴリ毎の単語頻度(ICF)という TF-DF-ICF 法によるカテゴリごとによる単語辞書の作成.

（２） 単語同士の記号列の完全マッチングによるその単語の出現の有無に基づいた頻度計算ではなく,単語同士の部分マッチングをアナログ値に持つ単語の圧縮率と定め,さらに閾値を適用し,単語頻度ヒストグラムを作成する方法.

（３） 従来研究が,全カテゴリについて共通した１つの辞書を用いて,カテゴリ毎の代表単語頻度ヒストグラムル作成のための単語頻度ヒストグラム群抽出と,識別対象の入力文書について単語頻度ヒストグラム抽出が行われている従来法に比べて,本研究では,１つの辞書を用いるのではなく,カテゴリごとの辞書を用いて文書カテゴリを代表する単語頻度ヒストグラムを作成し,また未知入力文書についてカテゴリごとの辞書を用いて単語頻度ヒストグラムを抽出し,それらの間の類似度計算に相関係数を利用.

上記の手法を 20newsgroups データセットを利用して提案法の有効性を検証する.この結果,識別が識別対象データについて,カテゴリの識別正解率の平均６８%を得ることが出来た.